Correlating morphosyntactic dialect variation with geographic distance: Local beats global

Jeszenszky, Péter ; Weibel, Robert

Correlating morphosyntactic dialect variation with geographic distance: Local beats global

Péter Jeszenszky, Robert Weibel

Department of Geography, University of Zurich (UZH),
Winterthurerstrasse 190, CH-8057, Zurich
Email: {peter.jeszenszky | robert.weibel}@geo.uzh.ch

1. Introduction

Similarly to Tobler's First Law of Geography, dialectology has its own postulate, termed the 'Fundamental Dialectological Postulate' (FDP): „Geographically proximate varieties tend to be more similar than distant ones" (Nerbonne & Kleiweg 2007: 154). This postulate seems intuitive, and thus several authors have tried confirming it by determining the degree of correlation between dialectal variation, expressed by a linguistic distance measure, and some geographic distance measure (e.g. Nerbonne & Kleiweg 2007; Spruit et al. 2009), all reporting (highly) significant correlations. While most authors have used Euclidean distance, some used travel time as a geographic distance measure that represents potential geographic language contact with an increased degree of realism (Gooskens 2004; Haynie 2012). However, in a recent study by Szmrecsanyi (2012) using corpus-based data about morphosyntax (i.e. grammatical constructs) in traditional English dialects, the FDP has been contested, reporting non-significant correlation.

The above studies are all rooted in linguistics, and have led to interesting results. From a geographical perspective, however, they all suffer from the crucial drawback of restricting the analysis to the —geographically speaking— global level, computing correlations for entire study areas, rather than exploring linguistic variation in more detail at the *local* level. Hence, they miss out on discovering regional differences in correlation structures, and on offering possible explanations of regionally different linguistic variation patterns. Also, global analysis alone will not be able to explain the large differences in the degrees of correlation reported in different studies.

Thus, the objective of our work is to enable the spatially differentiated comparison of linguistic variation and geographic distances, shedding new light on the FDP. For the case of morphosyntactic variation in Swiss German dialects, we present methods to establish global and local correlation between language and geographic distances, giving preliminary results and an outlook on possible extensions. While this work should be mainly beneficial for linguistics, we believe that it is also relevant to GIScience, since linguistic data represent a type of data that is uncommon in GIScience. Furthermore, we would like to show that dialectology and other strands of linguistics offer plenty of opportunities for GIScientists to contribute to advancing science at the interface between disciplines.

2. Data and Methods

2.1 Data

This study uses data from the Syntactic Atlas of German-speaking Switzerland (SADS; Bucheli & Glaser 2002). The SADS project was initiated in 2000 to map and study syntactical (i.e. grammatical) phenomena of Swiss German dialects. Close to 3,200 informants participated in a survey, providing answers to 118 questions, corresponding to

linguistic *variables*. Informants live in 383 municipalities, i.e. in approx. 25 % of the German speaking municipalities in Switzerland. An important feature of the SADS is that multiple informants occur per survey site, ranging between 3 and 26, with a median of 5 to 6 informants per site. Thus, linguistic variation, expressed by different *variants* for a given variable, exists also between respondents at each site. The following example shows this dual variation in a linguistic variable in the SADS:

English – 'I don't have enough change in order to buy a ticket.'
Standard German – 'Ich habe zu wenig Kleingeld um eine Fahrkarte zu lösen.'
Main variant 1. – 'Ich ha z wenig Münz **für** es Billet **z** lööse.'
Main variant 2. – 'Ich ha z wenig Münz **zum** es Billet **z** lööse.'

In this example, the linguistic *variable* is the syntax construct of the so-called infinitival complementizer, for which two *variants* exist, using 'für' and 'zum', respectively.

2.2 Methods

Linguistic (dis)similarity is often computed using edit distances, such as Hamming and Levenshtein distance (Spruit et al. 2009). However, since in the SADS multiple variants may occur per survey site, we had to use a different method. Figure 1, for two sample variables (Question I.01 and Question I.03) and two survey sites (Klosters, Flühli), shows the procedure of computing a linguistic distance — in this case, the *syntactic* distance — between a pair of sites.

Once the syntactic distances have been computed for all survey site pairs, the global correlation between the linguistic and the geographic distances between sites is computed. We use Pearson product-moment correlation and correlation established by the Mantel test.

Simply computing global correlations will not reveal the potential causes of linguistic variation, and is prone to ecological fallacy. This is improved in two ways. First, by focusing the analysis on a local subset of the study area. Second, by normalizing both the linguistic and geographic distances obtained, it becomes possible to compute residuals per site and thus analyze locally how well geographic distance predicts the observed linguistic distance.

Besides Euclidean distance, geographic distance was also represented by a travel time matrix provided by the Institute for Transport Planning and Systems at ETH Zurich (Fröhlich et al. 2004).

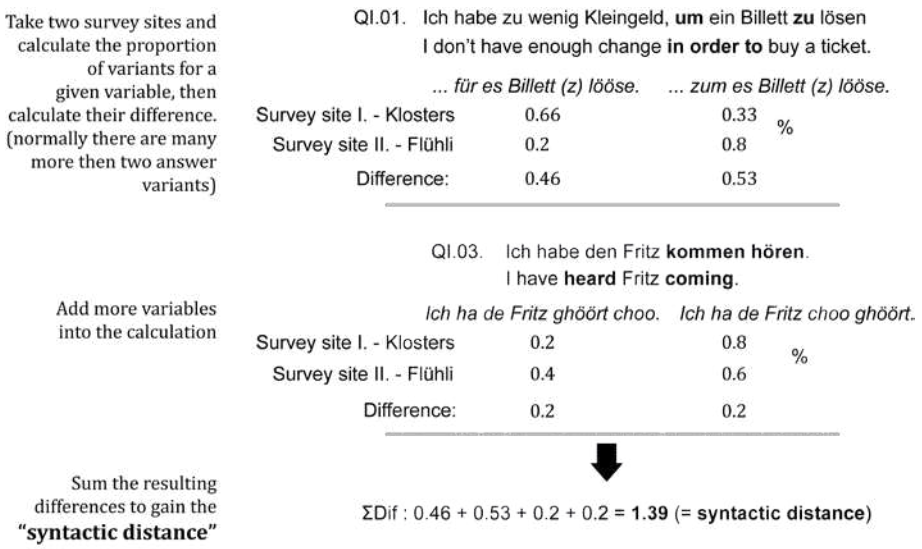

Figure 1: Workflow to compute the pairwise syntactic distance between two sites.

3. Results

So far, we have computed syntactic distances using 19 linguistic variables, which are hypothesized by the SADS linguists to be representative of the main morphosyntactic phenomena in Swiss German. Thus, the results reported below are preliminary from a dialectological perspective. However, they may nevertheless serve to illustrate the potential of our approach.

Tables 1 and 2 present the results of the correlation analysis on the global scale and for a particularly interesting local subset, the region between the Bernese Oberland and the German-speaking part of the Valais (BEOV, N = 45). All correlation coefficients are significant to highly significant (at least $p < 0.05$). As the right hand column of Tables 1 and 2 shows, the differences between the correlation coefficients at the global level as opposed to the coefficients at the BEOV level are significant, with the exception of the correlation the Mantel test results for both travel times. However, when comparing the correlations obtained with different distance measures, only very few were reported significant (results not shown in Tables 1 and 2). Only subtle differences between 0.722 and 0.747 exist for the global level and are thus not significant. In the BEOV subset, only one highly significant difference ($p < 0.01$) can be found between Euclidean distance and travel times 1950 in the Mantel test (0.366 vs. 0.750). Additionally, the difference between Euclidean distance and travel times 2000 in the Mantel test (0.366 vs. 0.707) is significant ($p < 0.05$). And one difference—between Euclidean distance and travel times 2000 in the Pearson correlation coefficients (0.307 vs. 0.578)—is almost significant ($p = 0.0582$).

The map in Figure 2 shows the survey sites, represented as Voronoi polygons to fill in the gaps between sites, colored according to their syntactic distance from a particular place, Schaffhausen, with the borders of the Swiss cantons overlaid. Normalizing the distances, residuals per site can be obtained, showing the degree of agreement between the two distance measures (Fig. 3). Thus, if the normalized syntactic distance from the survey site "Obersaxen" were in perfect linear agreement with the corresponding normalized Euclidean distance, no residuals would show in Figure 3. Figure 4 then maps the residuals of Figure 3 to geographic space. Finally, Figure 5 depicts the syntactic distances from "Adelboden" for the local subset BEOV in the area of the Bernese Oberland and the German speaking part of the Canton of Valais.

Table 1. Pearson correlation coefficients for global area and a regional subset.

For 19 variables	Syntactic distance (global, N = 383)	Syntactic distance (BEOV subset, N = 45)	Fisher's Z, one-tailed
Euclidean distance	0.722^{***}	0.307^{*}	***
Travel times by car - 1950	0.745^{***}	0.578^{***}	*
Travel times by car - 2000	0.743^{***}	0.524^{***}	*

* = P ≤ 0.05, ** = P ≤ 0.01, *** = P ≤ 0.001, ns = statistically not significant

Table 2. Mantel test results for global area and a regional subset.

For 19 variables	Syntactic distance (global, N = 383)	Syntactic distance (BEOV subset, N = 45)	Fisher's Z, one-tailed
Euclidean distance	0.747^{***}	0.366^{**}	***
Travel times by car - 1950	0.738^{***}	0.750^{***}	ns
Travel times by car - 2000	0.734^{***}	0.707^{***}	ns

* = P ≤ 0.05, ** = P ≤ 0.01, *** = P ≤ 0.001, ns = statistically not significant

4. Discussion

As Tables 1 and 2 show, all correlation coefficients are highly significant on the *global level*, independently of the correlation measure used. However, the difference between the results for the different geographic distance measures is not statistically significant.

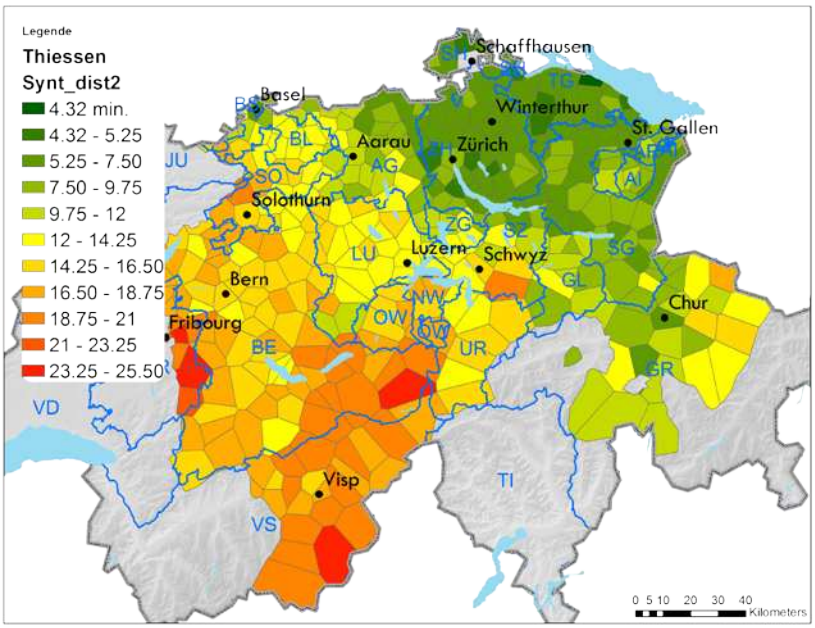

Figure 2: Syntactic distances from Schaffhausen.

The story is different at the *regional level*, represented by the BEOV subset. Here, we find generally lower correlations compared to the corresponding values at the global level, but we also find significant differences between the Euclidean and travel time distances. In the BEOV subset, a high mountain area is represented, where topography crucially influences travel times. Thus, travel time is a significantly better predictor at this more local level.

As Figure 2 shows for the example of Schaffhausen, the syntactic distances from this site exhibit a pattern that appears to largely follow the increase in Euclidean distance, with some exceptions. This suggests a possible explanation of the highly significant correlation with Euclidean distance on the global level, which at the same time does not differ significantly from correlation results obtained with travel times.

The differences between normalized syntactic and Euclidean distances (Fig. 3) follow a decreasing trend. They are positive at short ranges, meaning that the Euclidean distance underestimates short-range syntactic variation. The opposite is the case at long ranges, where Euclidean distance overestimates syntactic variation. This overestimation at long ranges makes sense, since geographic distance increases continuously, while the dialectal distance may only increase to a certain level. If two dialects become too dissimilar, they will be considered two different *languages*, as they are no longer mutually intelligible. This geographic pattern becomes even more apparent in the map of Figure 4.

Finally, Figure 5 shows some interesting patterns at the regional and local level for the BEOV subset, which represents high mountain topography, with secluded valleys. These patterns would not become apparent if the analysis was restricted to the global level. For instance, we could see a bridging effect of two mountain passes, the Gemmi Pass and the Grimsel Pass, respectively, which connect two sides of a high mountain range that largely exceeds 4,000 m.a.s.l. The Gemmi Pass being one of them, nowadays cannot be traversed by road but used to be a major pass in the Middle Ages when most dialect formation took place. Further work, however, is needed to explore these effects in more detail.

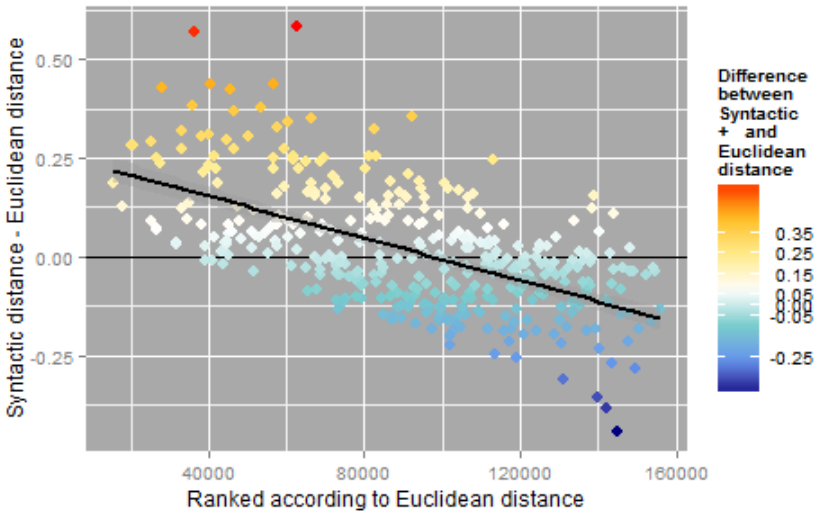

Figure 3: Residuals of syntactic distance and Euclidean distance for survey sites paired with the alpine village Obersaxen (cf. Figure 4).

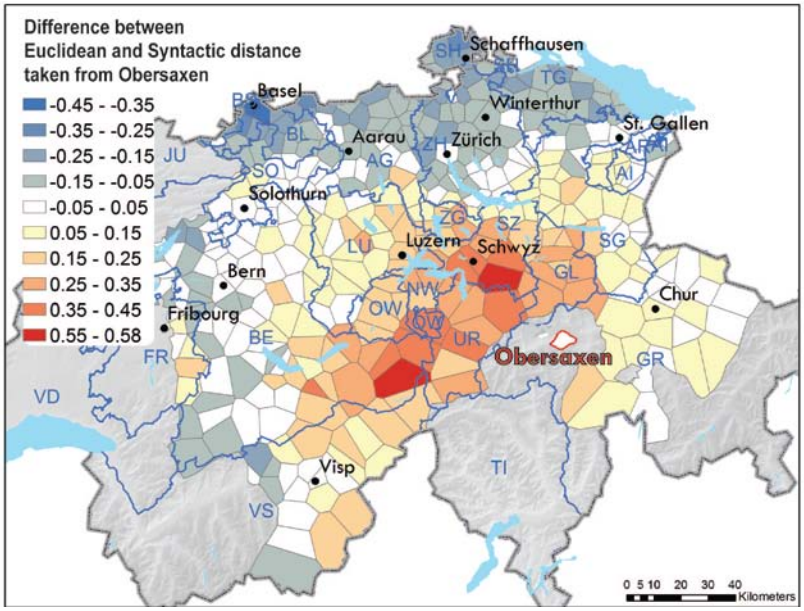

Figure 4: The residuals of Figure 3 mapped to geographic space.

5. Conclusions

We have shown how global correlation analysis with geographic distances in dialectology can be extended to the local level, painting a more differentiated picture of the dialectal variation across space. For the case of morphosyntactic variation represented by the SADS, we have been able to confirm the FDP, and show that different geographic distance measures only play out at the local level as a predictor variable.

Various extensions are possible. From a linguistic perspective, we will add more SADS variables and possibly also variables from other linguistic levels (lexis, phonetics, morphology). While today, travel times are increasingly approximating the concentric pattern of Euclidean distance, owing to ever improving accessibility, we will be extending the analysis to pre-1850 travel times, hypothesizing the results to differ significantly from those obtained with Euclidean distances. We will also explore other proxies of language contact such as linguistic gravity (Szmrecsanyi 2012), commuter matrices etc.

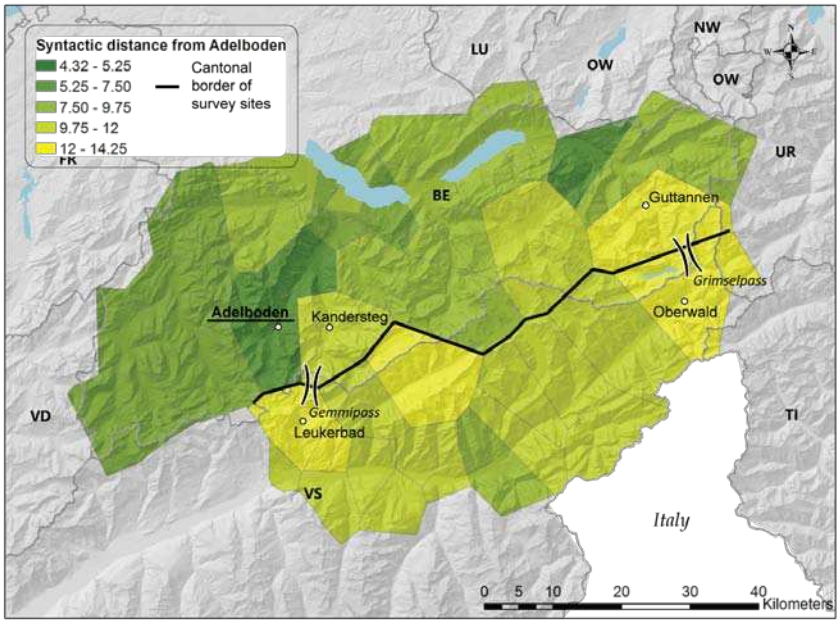

Figure 5: Map of syntactic distances from Adelboden in the BEOV subset. The cantonal border is formed by a major alpine drainage divide, bridged by two mountain passes.

From the methodological perspective, the current method of linear summation of syntactic distance assumes independence of variables, neglecting potential mutual correlation. Correlation analysis and dimension reduction could be explored. Finally, the most interesting extension will be to represent "geography" not only by geographic distances, but attempt to relate linguistic (i.e. syntactic) variation to geographical features, such as topographic, political or cultural borders.

Acknowledgements

This research represents part of the PhD project of the first author. Funding by the Swiss National Science Foundation through project SynMod (CR12I1-140716) is gratefully acknowledged. We are grateful to the Institute for Transport Planning and Systems within the Swiss Federal Institute of Technology for providing the travel time data, and for the Syntactic Atlas of German-speaking Switzerland (SADS) project for the syntactic data. Finally, we would like to thank Philipp Stöckle, German Department of UZH, for his valuable comments.

References

Bucheli, C., & Glaser, E. (2002). The Syntactic Atlas of Swiss German Dialects: Empirical and Methodological Problems. In S. Barbiers, L. Cornips, & S. van der Kleij (Eds.), *Syntactic Microvariation* (Vol. 2., pp. 41–73). Amsterdam: Meertens Institute Electronic Publications in Linguistics.

Fröhlich, P., Frey, T., Reubi, S., & Schiedt, H. U. (2004). *Entwicklung des Transitverkehrs Systems und deren Auswirkung auf die* Raumnutzung *in der Schweiz (COST 340): Verkehrsnetz-Datenbank* (No. 340) (p. 54).

Gooskens, C. (2004). Norwegian Dialect Distances Geographically Explained. In *Language Variation in Europe. Papers* from *the Second International Conference on Language Variation in Europe ICLAVE Vol. 2. 2004*. (p. 10).

Haynie, H. J. (2012). Studies *in the History and Geography of California Languages*. University of California, Berkeley.

Nerbonne, J., & Kleiweg, P. (2007). Toward a Dialectological Yardstick. *Journal of Quant. Ling.*, *14*(2), 148 p.

Spruit, M.R., Heeringa, W. & Nerbonne, J. (2009). Associations among Linguistic Levels. *Lingua*, 119(11), 1624–1642

Szmrecsanyi, B. (2012). Geography is overrated. In S. Hansen, C. Schwarz, P. Stoeckle, & T. Streck (Eds.), *Dialectological and Folk Dialectological Concepts of Space* (pp. 215–232). Berlin, Boston: De Gruyter.